平成２９年度学部学生による自主研究奨励事業研究成果報告書

ふりがな 氏　　名	いかり　しょうご 五十里 翔吾	学部 学科	基礎工学部	学年	2 年
ふりがな 共　同 研究者氏名	くろだ　かずき 黒田 和暉	学部 学科	基礎工学部	学年	2 年
	きみず　まさたか 木水 祐孝		基礎工学部		2 年
					年

アドバイザー教員 氏名	潮　俊光	所属	基礎工学研究科

研究課題名	強化学習を用いた PID 制御システム最適化に関する研究
研究成果の概要	研究目的、研究計画、研究方法、研究経過、研究成果等について記述すること。必要に応じて用紙を追加してもよい。（先行する研究を引用する場合は、「阪大生のためのアカデミックライティング入門」に従い、盗作剽窃にならないように引用部分を明示し文末に参考文献リストをつけること。）

1. はじめに

筆者の所属するコースの 2 年次選択授業である『知能システム学 PBL』の授業内で、筆者たちは試行錯誤しながら行動を最適化する理論的枠組み[1.1]である強化学習について実習を通して学んだ。そのさなか、ロボットの運動制御に用いられる PID ゲインを調整する過程が、まさしく「働きかけによって観測される結果が変化する対象に対し、不完全な知識から、最適な働きかけ方の系列を発見する問題」つまり強化学習問題[1.1]である事に気づき、上記の強化学習の方法がパラメータの最適化に応用できるのではないかと思い至った。本研究では、試行錯誤の末「勾配法」が適当な学習法であると結論づけ、1 関節ロボットアームの PID 制御について、シミュレーションと実機実験でその効果を検証した。今回確立した方法を用いると、限定された条件のもとでは、制御パラメータがいくつに増えてもシステム設計者が指定すべき値は、実質的に 1 つとなる。

2. 強化学習

2.1 Q-learning を用いた迷路の最適経路探索

まず、強化学習の考え方を学ぶため、迷路を学習により解くプログラムを実装した。

2.1.1 最適化アルゴリズム：価値反復法[2.1]

ベルマン方程式と呼ばれる再帰式を解くと、未来の収益期待値が計算できる。その解法の 1 つに、「Q-learning」がある。この方法では、行動価値関数を式(2.1)で更新し、最適値を得る。

$$Q(S_t, A_t) \leftarrow (1-\alpha)Q(S_t, A_t) + \alpha\left(R_t + \gamma \max_{a' \in A(s')} Q(S_{t+1}, a')\right) \quad -(2.1)$$

Qは行動価値、Aは行動、αは学習率、Rは報酬、γは未来の報酬の割引率である。

2.1.2 実装

上記のアルゴリズムに従って実装を行い、以下に示すように学習で最短経路が得られることを確認にした。（図 2-1）このとき、ゴール位置に遷移する試行のみから報酬が得られるように設定した。迷路左上の初期位置での行動価値が、図 2-2 のように増加していくことも確認できた。

2.2 倒立振子シミュレーターを学習により倒立させる制御器の制作

2.2.1 概要

研究の目標は、ロボットの運動制御に強化学習を応用することであるため、迷路課題の次は、よりその形に近い倒立振子を題材に選んだ。具体的には、OpenAI Gym で公開されている Cartpole-v0[2.2]を学習によって倒立させる課題に取り組み、学習を用いて解くプログラムを作成した。

2.2.2 シミュレータの仕様と制御アルゴリズム

Cartpole-v0 では、車の上にヒンジで固定されたポールを立たせ続けることが目標となる。シミュレータへの入力は左か右かの二種類である。（図 2-3）時間ステップtでの報酬は、試行がT_{end}で終了（転倒）した後、$T_{end}-t$で与える。そして価値反復法により離散化した角度、角速度での行動（左、右）に対する価値関数を計算し、それに基づいて倒立振子の制御を行う。

2.2.3 学習結果

最大の時間ステップを 200 とし、上記の方法で学習を進めた時の様子は図 2-4 のようになった。時間の経過とともに倒立時間が伸びている。

2.2.4 総括

2.1 で学んだ強化学習の手法を倒立振子という制御問題に応用することが出来た。PID 制御そのものについて理解を深めることで、さらなる応用が可能だと考えられる。

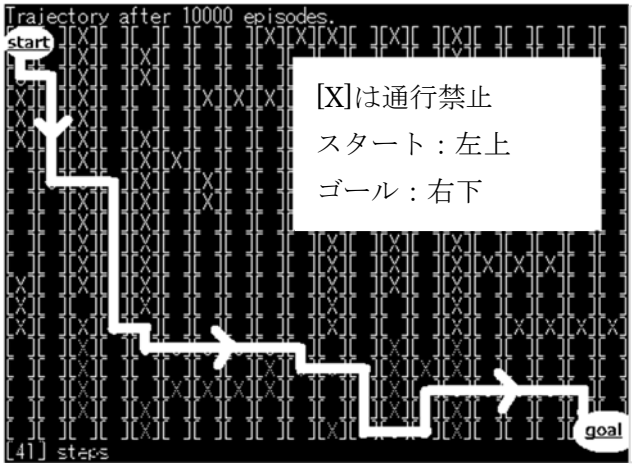

図 2-1

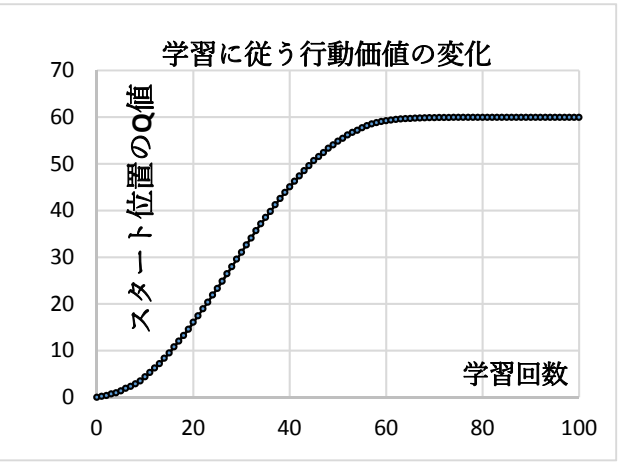

図 2-2

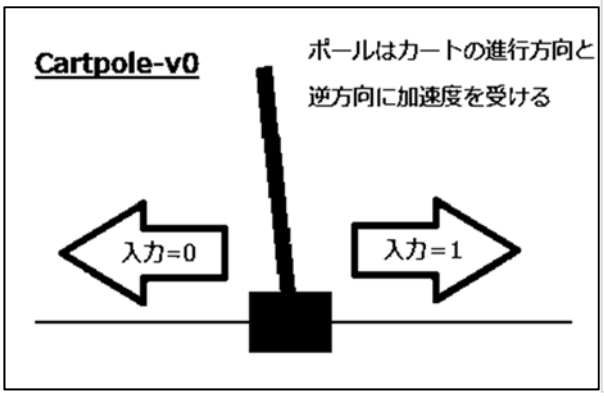

図 2-3

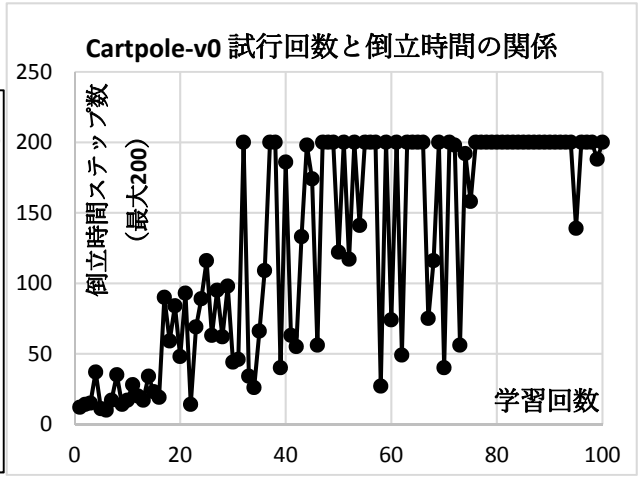

図 2-4

3. PID 制御に関する実験

3.1 目的

前述の通り、PID 制御ゲインを最適化することが目標であるため、その特徴について知る必要がある。また、実際に手動でゲインを調整してみることにより、学習を実装する際の参考となることが期待できる。

3.2 使用器具、測定方法

Vstone 社「ビュートバランサー2」は、PI 制御を用いて倒立振子の制御を実現している。本実験では、(3a)、(3b)に基づいて制御の質を数値化する実験を行った。

(3a) 十分時間経過後の特性の良さ

十分時間経過後のジャイロセンサの値の振幅をA_rとし、これが小さいほど良いとした。

(3b) 外乱に対する応答の良さ

大きさRの外乱に対する再静定までの時間をt_rとし、これらの比R/t_rを外乱に対する応答の良さとした。

これら両方について検討するため、式(3.1)に示すW（目的関数）をパラメータの良さと定義した。

$$W = 100 - \frac{A_r}{(R/t_r)\times 100} = 100 - t_r A_r/100R \quad -(3.1)$$

3.3 結果と考察

Wをそれぞれのゲインにおいて計算し、三次元棒グラフにプロットしたものが図 3-1 である。空白の部分は、外乱がない状態で倒立させることができなかったか、外乱後に再整定しなかった。縦軸がWである。このグラフから、1 つのゲインをもう一方を固定して段階的に変化させた時、パラメータの良さも段階的に変化することがいえる。このことから、$W = W(K_P, K_I, K_D)$とみなすことで、勾配法アルゴリズムに基づきゲインの良さを最大化できると推測される。

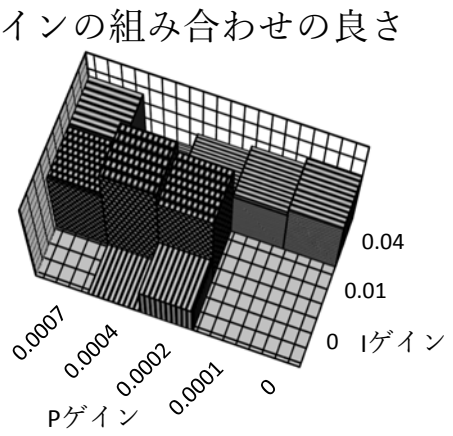

図 3-1

4. 方策勾配を参考にした学習方法による PID ゲイン最適化のシミュレーション実験

4.1 概要

3 節では、PID 制御ゲインの最適化には勾配法に基づいた学習法が有効との結論が得られた。本節では、勾配法の概要について説明し、シミュレーションにより有効性を検討する。

4.2 勾配法[4.1]

このアルゴリズムでは、方策をパラメータ$\boldsymbol{\theta}$によって表し、$\boldsymbol{\theta}$に関する勾配を用いて目的関数を最適化する。今回は、$\boldsymbol{\theta} = (K_P, K_I, K_D)$とし、以下のように勾配を近似して$\boldsymbol{\theta}$を更新する。

(4a) $\boldsymbol{\theta_1} = (K_P + \Delta p, K_I, K_D)$、$\boldsymbol{\theta_2} = (K_P - \Delta p, K_I, K_D)$としてそれぞれのゲインに対する目的関数の値を記録し、$\boldsymbol{\theta_1}$での報酬から$\boldsymbol{\theta_2}$での報酬を引いた値をJ_Pとして保存する。

(4b) J_Pについて、もし前回の試行での値との積が負になった場合はΔpを1.7^{-1}倍する。

(4c) 上の操作を他のゲインについても行い、$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + (J_P, J_I, J_D)$とする。

ただし、目的関数の値Rは、以下の式(4.1)で与える。

$$R=\begin{cases} C_1/\text{収束時間 (収束した場合)} \\ -C_1/\text{発散時間 (発散した場合)} \\ -C_2\times\text{試行終了時の偏差 (振動を続けた場合)} \end{cases} \quad (\mathrm{C}_1,\mathrm{C}_2\text{は定数}) \quad -(4.1)$$

4.3 学習のシミュレーション

4.3.1 制御量などの説明

シミュレーションでは、制御量の運動方程式を式(4.2)で与え、$x=0$の状態を目標値とした。

$$\mathrm{I}\ddot{x}+b\dot{x}-gx=0 \ (\text{ただし}I=0.1,b=50,g=9.8) \quad -(4.2)$$

実装にあたり文献[4.2]、[4.3]を参考にした。

4.3.2 結果と考察

$\boldsymbol{\theta}$の初期値を$\boldsymbol{\theta_0}=(5.0,10.0,0.0)$、探索のためにゲインに加減する値（ステップ幅）の初期値を$(\Delta p_0,\Delta i_0,\Delta d_0)=(1.0,1.0,1.0)$とし、260 回の学習を行った。その結果が図 4-1、及び図 4-2 である。

図 4-1 を見ると、エピソードを重ねるに従い、はじめは激しく振れ発散していたxが収束するようになっている。更に学習が進むとオーバシュートが小さくなり。収束までの時間も短くなっていることもわかる。このように、学習により制御系が改善される過程が確認できた。

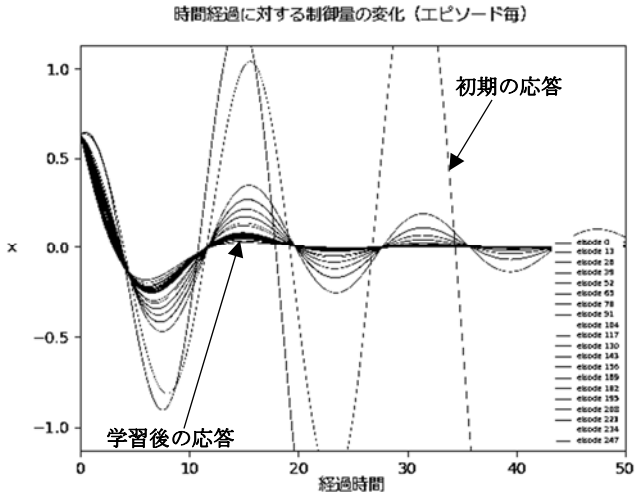

図 4-1

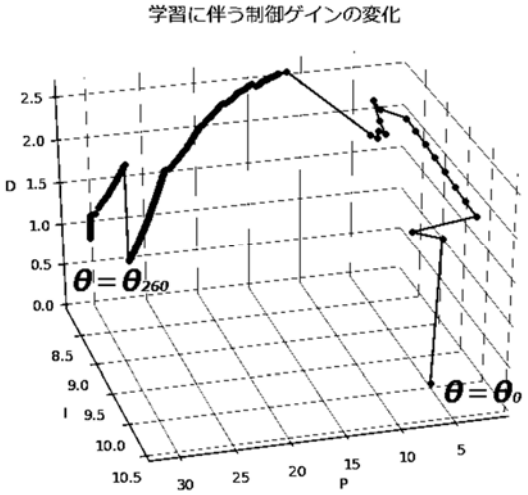

図 4-2

5. 実機を用いた学習の効果の検証

5.1 実験器具

図 5-1 が実験に用いたロボットアームの全景である。回転式のアームのモーターにはタミヤギヤードモーター540K30、制御用マイコンボードにはSTMircroelectronics 社の STM32F429I-DISC1、電源は Turnigy 社のリチウムポリマー電池を用いた。このアームを 90 度回転させるという課題に対する学習を計測する。

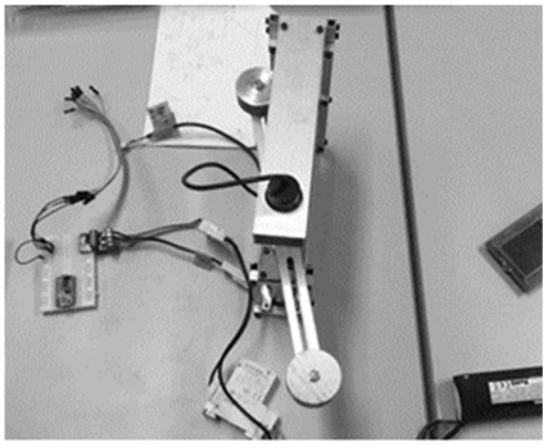

図5-1

5.2 実験結果

図 5-2 及び図 5-3 に、定められた初期値のもとステップ幅を(0.08,0.3,0.05)として学習させた結果を示す。図 5-2 の初期値は図に示す通りであり、図 5-3 の初期値は(0.1,0.7,3.0)である。

5.3 考察と今後の展望

図 5-2 からは収束時間の短縮、図 5-3 からは過渡応答の改善がそれぞれ確認できる。すなわち、勾配法に基づく学習法の効果が限定的な条件のもとで検証できた。すなわち初期ゲイン、初期ステップ幅により加減算されたゲインの両方で応答が収束する場合である。使用したモーターの耐熱性の問題で、4 節のシミュレーションでは示すことができた、初期ゲインで応答が発散する場合の効果は実験することができなかった。実験時に学習を 5 回で終了させているのも、同様の理由による。この問題を解決するためには、より少ない探索回数で勾配を求められるようにすればよい。その方法としては、過去の試行データから目的関数のおよその形状を推定することが考えられる。このように学習過程を洗練することで、より複雑な系に対しても応用が可能となるであろう。

今回用いた方法では、設計者が決めるべきパラメータは式(4.1)のC_1とC_2であった。簡単のため、実験では$C_1 = 4.0, C_2 = 20.0$とした。しかし、実際にはC_1とC_2は時間と偏差のスケール合わせのために異なる値となっているため、計算によりその適切な比を求めることが可能である。すなわち、発散の条件を$|x| \geq A_M$とし、試行を$T_{end}[s]$で終了させるとすると、振動と発散の境界値での目的関数値が等しくなればよく、条件$\mathrm{C}_2 \times A_M = \mathrm{C}_1/T_{end}$を得ることができる。この計算により

$$\frac{C_1}{C_2} = \frac{A_M}{T_{end}}$$

となるように定数を設定することで、制御パラメータの数によらず、設計者は 1 つの定数のみ検討すればよいことになる。

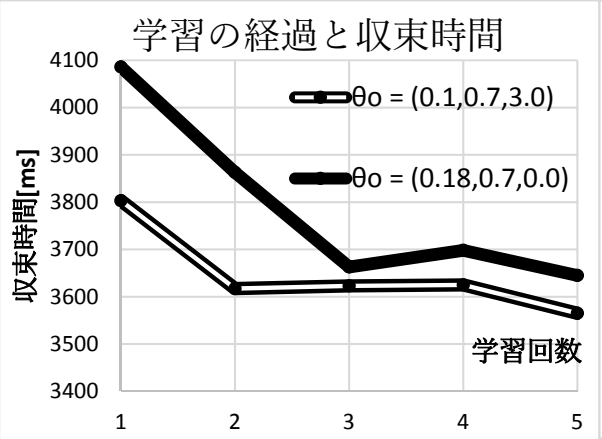

図 5-2

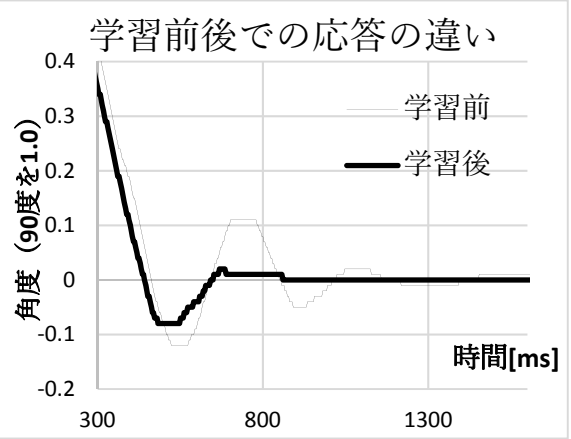

図 5-3

6. 参考文献

[1.1] 『これからの強化学習』(森北出版) 牧野貴樹 1.1 強化学習とは
[2.1] 同上 澁谷長史、牧野貴樹 1.3 価値反復に基づくアルゴリズム
[2.2] OpenAI Gym
https://gym.openai.com/envs/CartPole-v0/（2017 年 12 月 3 日閲覧）
[4.1] 『ゼロから作る deep learning』(オライリー・ジャパン) 斎藤康毅
[4.2] scipy で 2 階常微分方程式の数値解を求める
https://qiita.com/chase0213/items/95a107c013e4a6dbd7b5（2017 年 12 月 3 日閲覧）
[4.3] python でシミュレータを作ろう
https://chive.red-queen.moe/?p=30（2017 年 12 月 3 日閲覧）